10/520615
PCT/AU2003/000904 Rec'd PCT/PTO 10 JAN 2005
Received 11 October 2004

A LINK GENERATION SYSTEM

FIELD OF THE INVENTION

The present invention relates to a link generation system and process for generating links for a structured data set or network site such as a web site.

BACKGROUND

The ever increasing amount of information available on the Internet can make it extremely difficult to locate information relevant to a topic of interest. In the case of information available on the world-wide web, search engines have been developed for generating lists of hypertext markup language (HTML) documents or web pages matching one or more search terms supplied by a user. These lists of pages are generated from inverted indices generated by analysing the content of individual web pages. These web pages are retrieved by software modules known as spiders or web-crawling agents that crawl the web, using the hypertext transfer protocol (HTTP) to retrieve individual web pages, analyse content of those pages, and generate indices. This may involve identifying hyperlinks to other web pages, retrieving those linked pages, and analysing their content. Spiders can be used to generate indices for the world-wide web itself, or can be restricted to one or more specified web sites.

A web site can be viewed as a directed graph or digraph, with the servable content forming the nodes in the graph and directed links between the nodes corresponding to hypertext links within web pages of the site. A spider begins at one of the nodes in a web site, and then follows the links from that node to other nodes, and so on. The spider can perform whatever processing is desired for the nodes as it encounters them. In the case of a search engine spider, this involves indexing node content, but other spider types can be used to perform other tasks such as checking for broken hyperlinks or spell checking documents.

Unfortunately, not all web sites are completely connected - many have pages that are not directly connected to the rest of the web site through a hypertext link. In such a disconnected web site, a spider is unable to visit all of the nodes of the web site. This problem is especially pronounced in sites whose web pages include dynamic content. In the case of an indexing spider, a significant proportion of a site's content may not be accessible by a corresponding search engine. As more web sites convert their content from pre-existing, static web pages to more flexible and easier to maintain web pages including dynamically generated content, this problem will become even more significant.

Lack of full connectedness in a web site is also a potential problem for web site administrators who are trying to track their site's content. Without a completely connected graph of the site, it can be a difficult task to find all of the site content. For large sites with many content contributors, this task can become almost impossible.

Content that is not indexed by search engines has been referred to as 'the invisible web,' because it is not generally visible. It has even been suggested that the majority of information available on the web is invisible. Because invisible content is inaccessible to search engines, it decreases the visibility of web sites with invisible content, and degrades the usefulness of the web in general by making such content difficult to find.

It is desired, therefore, to provide a link generation system and process that alleviate one or more of the above difficulties, or at least to provide a useful alternative to existing link generation systems and processes.

SUMMARY OF THE INVENTION

In accordance with the present invention there is provided a link generation process executed by a computer system, including:

processing data files of a network site to identify valid parameters for generating dynamically generated content of said network site; and

generating encoded links for accessing said dynamically generated content, said encoded links including said parameters and being in a form suitable for an indexing agent to allow indexing of said dynamically generated content.

The present invention also provides a link generation process executed by a computer system, including generating at least one encoded link for retrieving dynamic content data of a hierarchical data set in response to selecting said at least one encoded link, said at least one encoded link including one or more parameters for generating said dynamic content data and being in a form suitable for an indexing agent to allow indexing of said dynamic content data.

The present invention also provides a link generation process, including:

generating encoded links for dynamic content of a network site, each of said encoded links including one or more parameters for use in generating said dynamic content and being in a form suitable for an indexing agent to allow indexing of said dynamic content;

receiving requests from an indexing agent for content of said site; and

responding to said requests with said encoded links and said dynamic content corresponding thereto for indexing.

The present invention also provides a link generation system, including:

a content discovery module for processing data files of a network site to identify servable data and parameters for generating dynamically generated content of said servable data; and

a link generator for generating links to said servable data to allow said servable content to be accessed using said links, said links including encoded links for accessing said dynamically generated content, said encoded links including said parameters and being in a form suitable for an indexing agent to allow indexing of said dynamically generated content.

The present invention also provides a link generation system, including:

one or more content discovery modules for processing data files of respective network servers to identify servable data and parameters for generating dynamically generated content of said servable data; and

a link generator for generating links to said servable data to allow said servable content to be accessed using said links, said links including encoded links for accessing said dynamically generated content, said encoded links including said parameters and being in a form suitable for an indexing agent to allow indexing of said dynamically generated content.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred embodiments of the present invention are hereinafter described, by way of example only, with reference to the accompanying drawings, wherein:

Figure 1 is a block diagram of a preferred embodiment of a link generation system connected to a remote user agent via a communications network;

Figure 2 is a flow diagram of a table of contents (TOC) generation process executed by the link generation system;

Figure 3 is a flow diagram of a table of contents selection process executed by the link generation system;

Figure 4 is a flow diagram of a directory TOC generation process of the TOC generation process;

Figure 5 is a flow diagram of a script TOC generation process of the TOC generation process;

Figure 6 is a flow diagram of a dynamic page parameter TOC generation process of the TOC generation process;

Figure 7 is a flow diagram of a dynamic page link generation process executed by the link generation system;

Figure 8 is a block diagram of a second preferred embodiment of the link generation system;

Figure 9 is a block diagram of a third preferred embodiment of the link generation system; and

Figure 10 is a block diagram of a fourth preferred embodiment of the link generation system.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

As shown in Figure 1, a link generation system 100 includes a content discovery module 102, a web server map (WSM) database 104, a table of contents (TOC) generation module 106, a dynamic page proxy module 108, a TOC server 110, a servlet 118, a web server 112, a scripting language module 120, a database module 122, a content database 124, static content files 126 (*e.g.*, hypertext markup language (HTML) and image files), and scripts 128 for generating dynamic content. The link generation system 100 can be

10/520615
PCT/Rec'd PCT/PTO 10 JAN 2005
PCT/AU2003/000904
Received 11 October 2004

CLAIMS:

1. A link generation process executed by a computer system, including:

processing data files of a network site to identify valid parameters for generating dynamically generated content of said network site; and

generating encoded links for accessing said dynamically generated content, said encoded links including said parameters and being in a form suitable for an indexing agent to allow indexing of said dynamically generated content.

2. A process as claimed in claim 1, wherein said generating includes generating a table of contents for content of said network site, said table of contents including said encoded links.

3. A process as claimed in claim 2, wherein said table of contents includes links to static content of said network site.

4. A process as claimed in claim 2, wherein said table of contents includes one or more pages, at least one of said pages including one or more links to content of said network site.

5. A process as claimed in claim 2, wherein said table of contents includes a plurality of pages, each of said pages including one or more links to respective others of said pages, at least one of said pages including one or more links to content of said network site.

6. A process as claimed in claim 5, wherein links in said table of contents pages are arranged as a hierarchy corresponding to content of said network site.

7. A process as claimed in claim 2, including generating a link to a table of contents page for a script for dynamically generating content on the basis of supplied parameters,

wherein said table of contents page for said script includes a plurality of encoded links corresponding to respective parameters for said script.

8. A process as claimed in claim 7, wherein the table of contents page for said script includes a plurality of encoded links corresponding to respective combinations of parameters and parameter values for said script.

9. A process as claimed in claim 7, wherein the table of contents page for said script includes at least one link to a further table of contents page including links corresponding to respective parameters or parameter values for said script and including at least one common parameter or parameter value

10. A process as claimed in claim 1, wherein said data files include at least one of web server configuration files, scripts, and database tables.

11. A process as claimed in claim 1, wherein said processing includes processing scripts of said network site to identify valid database query parameters on the basis of structured query language statements of said scripts.

12. A process as claimed in claim 1, wherein said processing includes processing said data files to identify valid combinations of database query parameters and values for generating said dynamically generated content.

13. A process as claimed in claim 12, wherein said processing includes processing database tables associated with said network site to identify said valid combinations of database query parameters and values.

14. A process as claimed in claim 1, wherein said encoded links are encoded as links to static content.

15. A process as claimed in claim 1, wherein each of said encoded links includes a suffix that indicates a type of dynamically generated content for the link.

16. A process as claimed in claim 1, wherein said encoded links include at least one link having a prefix identifying the link as a link to a table of contents page, and at least one link having a prefix identifying the link as a link to content of said network site.

17. A process as claimed in claim 1, including:

receiving a request for content of said network site from a remote agent;

determining whether said remote agent is an indexing agent;

sending a table of contents page to said remote agent if said remote agent is an indexing agent; and

sending the requested content to said remote agent if said remote agent is not an indexing agent.

18. A process as claimed in claim 1, wherein said encoded links are also URI-encoded.

19. A process as claimed in claim 1, wherein said processing includes processing said data files to identify all servable static content of said network site and all servable dynamically generated content of said network site; and wherein said generating includes generating links to said servable static content and said servable dynamically generated content to provide a table of contents for all servable content of said network site.

20. A process as claimed in claim 1, including processing scripts of said network site to determine request data for retrieving said dynamically generated data; wherein said encoded links are generated on the basis of said request data and said parameters.

21. A process as claimed in claim 20, wherein said step of processing scripts includes processing said scripts to determine access data for accessing a database of said network site to generate said dynamically generated content.

22. A process as claimed in claim 1, wherein said steps of processing and generating are executed at periodic intervals.

23. A process as claimed in claim 1, wherein said steps of processing and generating are executed in response to receiving a request for content of said network site.

24. A process as claimed in claim 1, including receiving a request generated in response to selecting one of said encoded links, translating said request, and forwarding the translated request to said network site to access corresponding dynamically generated content of said network site.

25. A process as claimed in claim 24, wherein said translated request is an HTTP GET request.

26. A process as claimed in claim 24, wherein said translated request is an HTTP POST request.

27. A process as claimed in claim 1, including sending said encoded links to a remote indexing agent to allow said dynamically generated data to be indexed.

28. A process as claimed in claim 1, including sending said encoded links to a remote system using one of HTTP PUT, HTTP POST, FTP, and SMTP.

29. A process as claimed in claim 5, wherein all servable data of said network site can be accessed via selection of any one of the links to said pages.

30 A process as claimed in claim 1, wherein said table of contents is generated in one of HTML, XML, HCL, and sitelist txt formats

31 A link generation process executed by a computer system, including generating at least one encoded link for retrieving dynamic content data of a hierarchical data set in response to selecting said at least one encoded link, said at least one encoded link including one or more parameters for generating said dynamic content data and being in a form suitable for an indexing agent to allow indexing of said dynamic content data.

32. A link generation process as claimed in claim 31, including generating a list of links to content data of at least one node of said hierarchical data set, said links including said at least one encoded link.

33. A link generation process as claimed in claim 32, wherein said generating includes generating links to all available data of said hierarchical data set.

34. A link generation process as claimed in claim 32, wherein said links include one or more direct links to content data of said hierarchical data set.

35. A link generation process as claimed in claim 32, wherein said links include one or more indirect links to content data of said hierarchical data set.

36. A link generation process as claimed in claim 32, wherein said links include at least one of a direct and an indirect link to content data of said node.

37. A link generation process as claimed in claim 32, wherein said list of links corresponds to a node of said hierarchical data set.

38. A link generation process as claimed in claim 32, wherein said hierarchical data set includes at least one web site.

39. A link generation process as claimed in claim 31, wherein said at least one encoded link includes at least one encoded POST query.

40. A link generation process as claimed in claim 31, wherein said at least one encoded link includes at least one encoded GET query

41. A link generation process, including:

generating encoded links for dynamic content of a network site, each of said encoded links including one or more parameters for use in generating said dynamic content and being in a form suitable for an indexing agent to allow indexing of said dynamic content;

receiving requests from an indexing agent for content of said site; and

responding to said requests with said encoded links and said dynamic content corresponding thereto for indexing.

42. A process as claimed in claim 41, wherein said links are generated as one or more of hyperlinks, XML elements, and text.

43. A link generation system having components for executing the steps of any one of claims 1 to 42.

44. A computer readable storage medium having stored thereon program code for executing the steps of any one of claims 1 to 42.

45. A link generation system, including:

a content discovery module for processing data files of a network site to identify servable data and parameters for generating dynamically generated content of said servable data; and

a link generator for generating links to said servable data to allow said servable content to be accessed using said links, said links including encoded links for accessing said dynamically generated content, said encoded links including said parameters and being in a form suitable for an indexing agent to allow indexing of said dynamically generated content.

46. A link generation system as claimed in claim 45, wherein said link generator is adapted to process a database of said network site to determine said parameters

47. A link generation system as claimed in claim 45, including a proxy server for receiving a request generated in response to selecting one of said encoded links, translating said request, and forwarding the translated request to said network site to access corresponding dynamically generated data of said network site.

48. A link generation system, including:

one or more content discovery modules for processing data files of respective network servers to identify servable data and parameters for generating dynamically generated content of said servable data; and

a link generator for generating links to said servable data to allow said servable content to be accessed using said links, said links including encoded links for accessing said dynamically generated content, said encoded links including said parameters and being in a form suitable for an indexing agent to allow indexing of said dynamically generated content.